Argonne Leadership Computing Facility: Mira Preparation and Recent Application Advances

Raymond Loy

Applications Performance Engineering

Argonne Leadership Computing Facility

Special thanks to Jeff Hammond, William Scullin, William Allcock, Kalyan Kumaran, and David Martin

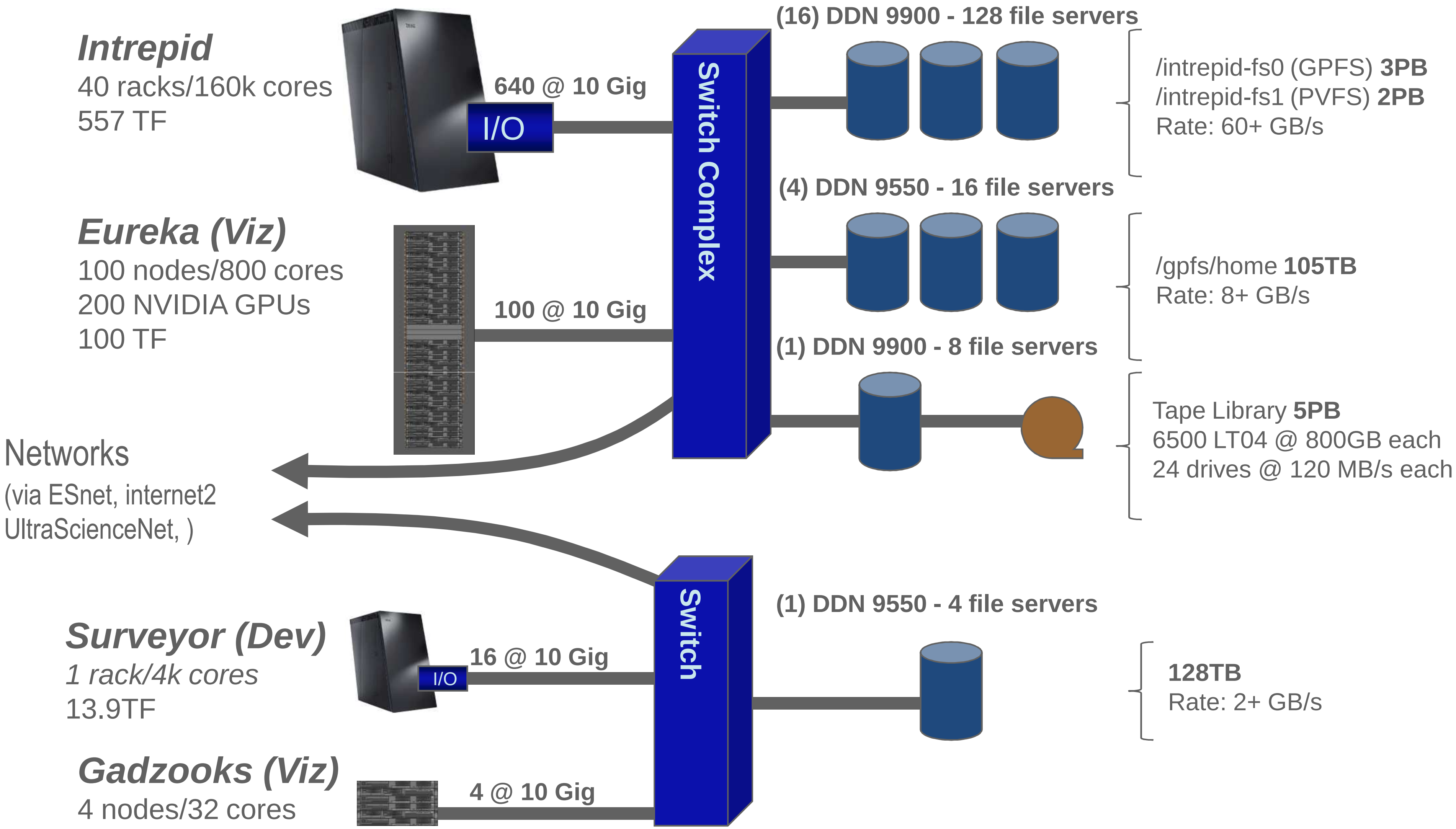
Argonne Leadership Computing Facility

- ALCF was established in 2006 at Argonne to provide the computational science community with a leading-edge computing capability dedicated to breakthrough science and engineering
- One of two DOE national Leadership Computing Facilities (the other is the National Center for Computational Sciences at Oak Ridge National Laboratory)
- Supports the primary mission of DOE's Office of Science Advanced Scientific Computing Research (ASCR) program to discover, develop, and deploy the computational and networking tools that enable researchers in the scientific disciplines to analyze, model, simulate, and predict complex phenomena important to DOE.
- Intrepid Allocated: 60% INCITE, 30% ALCC, 10% Discretionary

Argonne Leadership Computing Facility

- *Intrepid* - ALCF Blue Gene/P System:
  - 40,960 nodes / 163,840 PPC cores
  - 80 Terabytes of memory
  - Peak flop rate: 557 Teraflops
  - Linpack flop rate: 450.3
  - #13 on the Top500 (Nov 2010)
- *Eureka* - ALCF Visualization System:
  - 100 nodes / 800 2.0 GHz Xeon cores
  - 3.2 Terabytes of memory
  - 200 NVIDIA FX5600 GPUs
  - Peak flop rate: 100 Teraflops
- Storage:
  - 6+ Petabytes of disk storage with an I/O rate of 80 GB/s
  - 5+ Petabytes of archival storage (10,000 volume tape archive)

ALCF Resources - Overview



DOE INCITE Program

Innovative and Novel Computational Impact on Theory and Experiment

- **Solicits large computationally intensive research projects**
  - To enable high-impact scientific advances
  - Call for proposal opened once per year (2012 call closes 6/30/2011)
  - INCITE Program web site: http://hpc.science.doe.gov/
- **Open to all scientific researchers and organizations**
  - Scientific Discipline Peer Review
  - Computational Readiness Review
- **Provides large computer time & data storage allocations**
  - To a small number of projects for 1-3 years
  - Academic, Federal Lab and Industry, with DOE or other support
- **Primary vehicle for selecting principal science projects for the Leadership Computing Facilities** (*60% of time at Leadership Facilities*)
- **In 2010, 35 INCITE projects allocated more than 600M CPU hours at the ALCF**

DOE ALCC Program

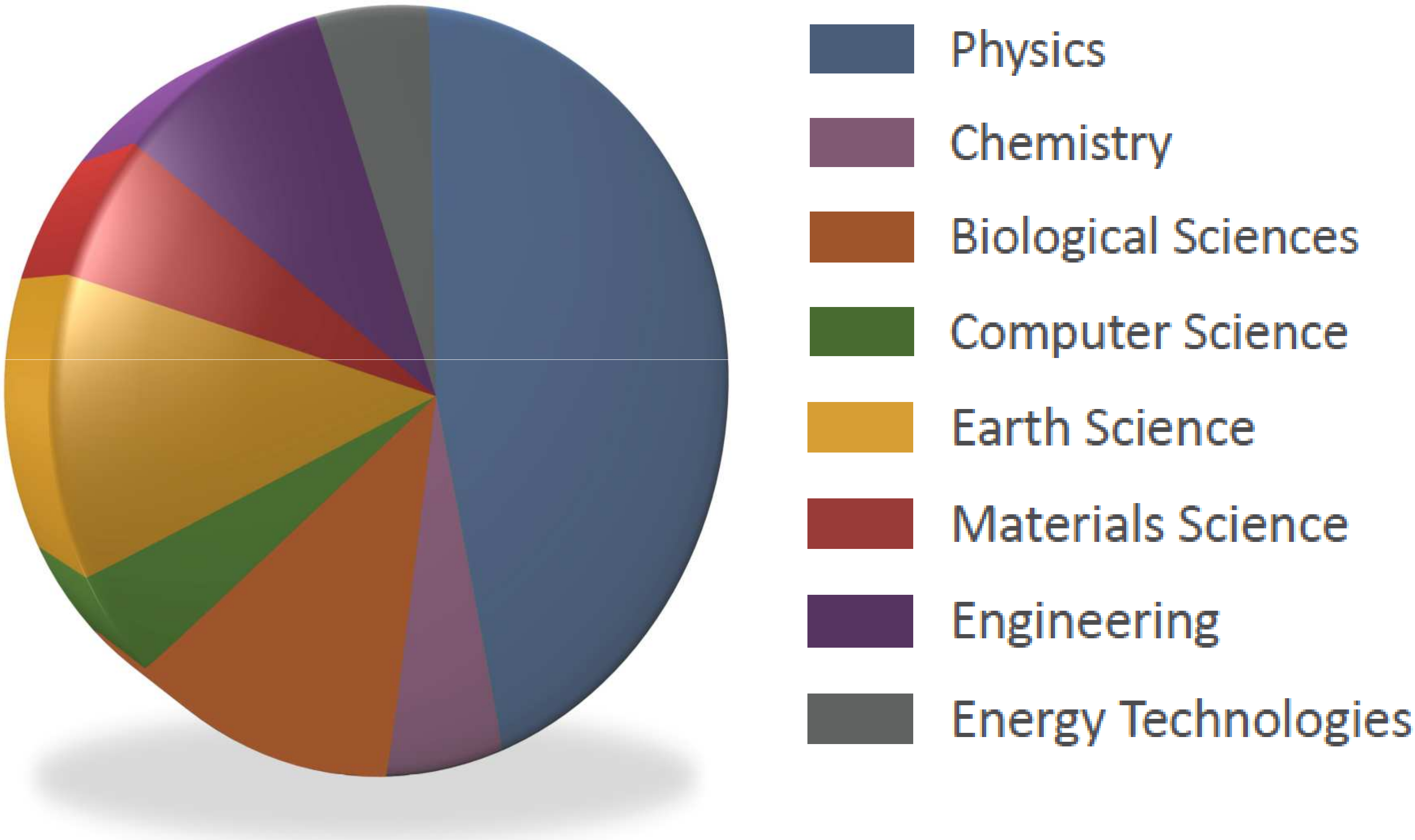
ASCR Leadership Computing Challenge

- Allocations for projects of special interest to DOE with an emphasis on high risk, high payoff simulations in areas of interest to the department's energy mission (30% of the core hours at Leadership Facilities)
- Awards
  - Last round granted in June, 2010
  - Call for 2011 allocations closed Feb 15, 2011
- http://science.energy.gov/ascr/facilities/alcc/
- 10 awards at ALCF in 2010 for 300+ million core hours

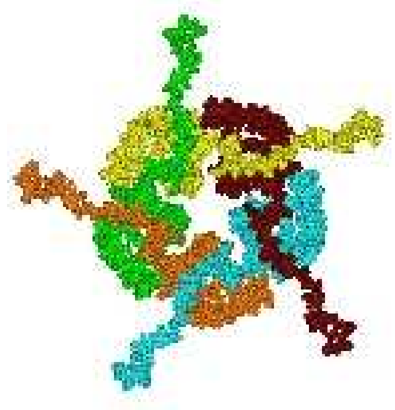
Discretionary Allocations

- Time is available for projects without INCITE or ALCC allocations!
- ALCF Discretionary allocations provide time for:
  - Porting, scaling, and tuning applications
  - Benchmarking codes and preparing INCITE proposals
  - Preliminary science runs prior to an INCITE award
  - Early Science Program
- To apply go to the ALCF allocations page
  - www.alcf.anl.gov/support/gettingstarted

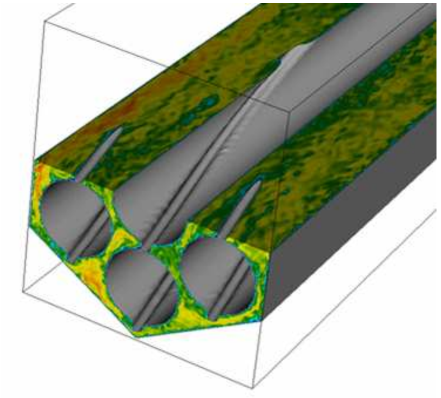
Percentage of compute hours used by scientific discipline
January–October 2010
Total hours used: 849 Million



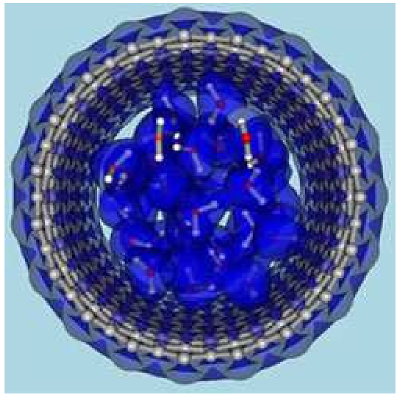
ALCF Projects Span Many Domains



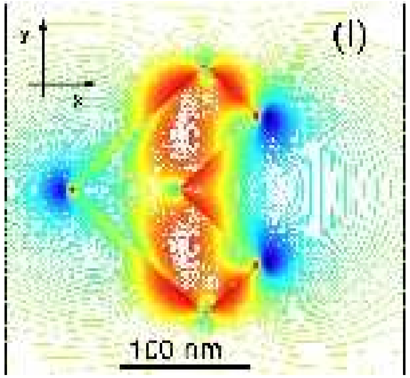
Life Sciences
U CA-San Diego



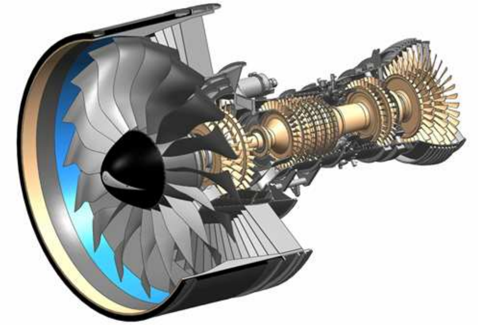
Applied Math
Argonne Nat'l Lab



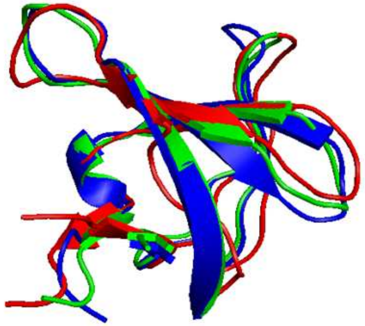
Physical Chemistry
U CA-Davis



Nanoscience
Northwestern U



Engineering Physics
Pratt & Whitney



Biology
U Washington

ALCF Timeline

2004

– DOE-SC selected the ORNL, ANL and PNNL team for Leadership Computing Facility award

2005

– Installed 5 teraflops Blue Gene/L for evaluation

2006

– Began production support of 6 INCITE projects, with BGW
– Continued code development and evaluation
– “Lehman” Peer Review of ALCF campaign plans

2007

– Increased to 9 INCITE projects; continued development projects
– Installed 100 teraflops BlueGene/P (late 2007)

2008

– Began support of 20 INCITE projects on BG/P
– Added 557 Teraflops BG/P

2009

– 28 Projects / 400 M CPU-hours

2010

– 35 Projects / 656 M CPU-hours

The Next Generation ALCF System: BG/Q

- DOE has approved our acquisition of “Mira”, a 10 Petaflops Blue Gene/Q system. An evolution of the Blue Gene architecture with:
  - 16 cores/node
  - 1 GB of memory per core, nearly a TB of memory in aggregate
  - 48 racks (over 780k cores)
  - 384 I/O nodes (128:1 Compute:I/O)
  - 32 I/O nodes for logins and/or data movers
  - Additional non-I/O login nodes
  - 2 service nodes
  - IB data network; 70 PB of disk with 470 GB/s of I/O bandwidth
  - Power efficient, water cooled
- Argonne and Livermore worked closely with IBM over the last few years to help develop the specifications for this next generation Blue Gene system
- 16 Projects Accepted into the Early Science Program
- Applications running on the BG/P should run immediately on the BG/Q, but may see better performance by exposing greater levels of parallelism at the node level

ALCF-2: Blue Gene/Q (Mira)
The story so far

Jan 2009

- CD0 approved

Jul 2009

- Leman Review (CD1/2a) passed

Jul 2010

- Lehman Review (CD2b/3) passed

Aug 2010

- Contract approved

2011

- BG/Q Early Science Program begins

ALCF-2: Blue Gene/Q (Mira)
What's next?

Mid 2011

- Early Access System
  - Approximately 128 nodes + 1 I/O node
  - Located at IBM, leased for ALCF use

Spring 2012

- T&D System delivery
  - 1-2 racks , 128:1 compute:IO node ratio (Same as Mira)

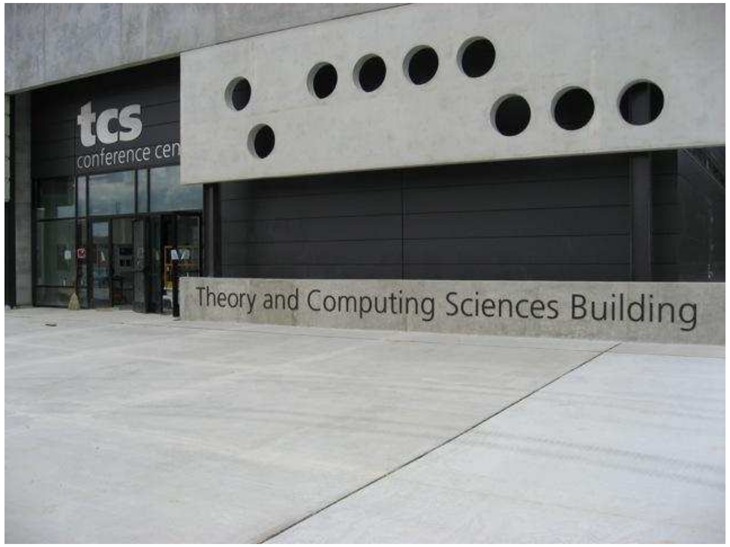
2012

- Mira delivery expected

2013

- Mira acceptance

TCS: *Future Home of Mira*



- 7 stories
- 25,000 ft^2 computing center
- 18,000 ft^2 library
- 10,000 ft^2 advanced digital laboratory
- 7,000 ft^2 conference center
- 30 conference rooms
- 3 computational labs
- 700 employees from 6 divisions

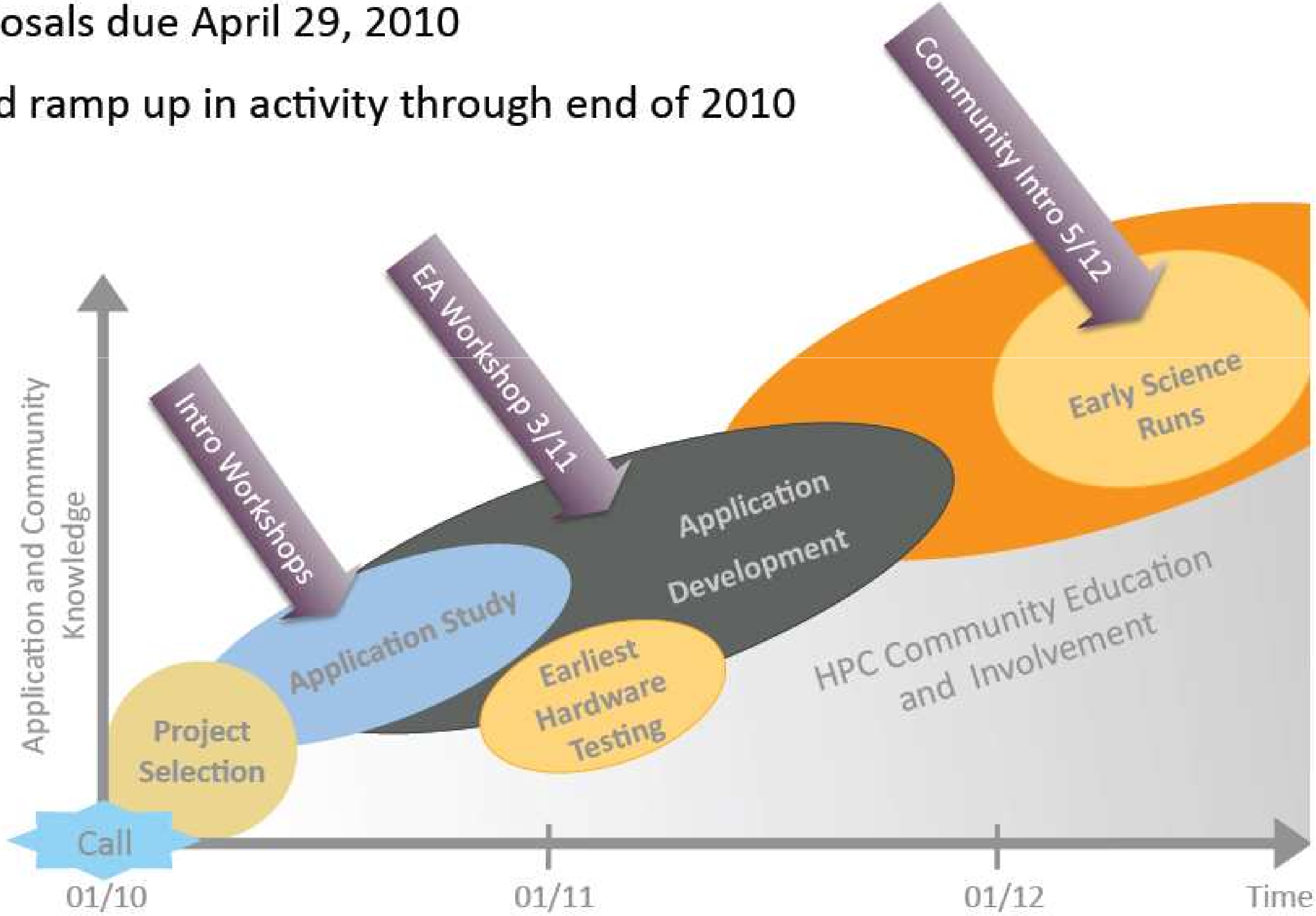
Preparing for Mira - Chilled Water Plant

Early Science Program

- "In early 2012 the ALCF will be installing at least 10PF of a next- generation Blue Gene.   We are asking the community to help us make this deployment as successful and productive as possible."
- Goals
  - Help us shake-out the system and software stack using real applications
  - Develops community and ALCF expertise on the system
  - A stable and well- documented system moving into production
  - Exemplar applications over a broad range of fields
  - At least 2 billion core-hours to science
- 2010 ESP Proposal Timeline
  - January 29th - Call for Proposals Issued
  - April 29th – Call for Proposals Closed
  - August – ESP Awards Announced
  - October 18-19 - Early Science Program Kickoff Workshop
  - Post docs start

Early Science Program Timeline

- Call Opens January 2010
- Proposals due April 29, 2010
- Rapid ramp up in activity through end of 2010



Early Science Projects

- **Climate-Weather Modeling Studies Using a Prototype Global Cloud-System Resolving Model**
  - PI: Venkatramani Balaji (Geophysical Fluid Dynamics Laboratory)
- **Materials Design and Discovery: Catalysis and Energy Storage**
  - PI: Larry A. Curtiss (Argonne National Lab)
- **Direct Numerical Simulation of Autoignition in a Jet in a Cross-Flow**
  - PI: Christos Frouzakis (Swiss Federal Institute of Technology)
- **High Accuracy Predictions of the Bulk Properties of Water**
  - PI: Mark Gordon (Iowa State University)
- **Cosmic Structure Probes of the Dark Universe**
  - PI: Salman Habib (Los Alamos National Laboratory)
- **Accurate Numerical Simulations Of Chemical Phenomena Involved in Energy Production and Storage with MADNESS and MPQC**
  - PI: Robert Harrison (Oak Ridge National Lab)

Early Science Projects (con't)

- **Petascale, Adaptive CFD**
  - PI: Kenneth Jansen (University of Colorado – Boulder)
- **Using Multi-scale Dynamic Rupture Models to Improve Ground Motion Estimates**
  - PI: Thomas Jordan (University of Southern California)
- **High-Speed Combustion and Detonation (HSCD)**
  - PI: Alexei Khokhlov (University of Chicago)
- **Petascale Simulations of Turbulent Nuclear Combustion**
  - PI: Don Lamb (University of Chicago)
- **Lattice Quantum Chromodynamics**
  - PI: Paul Mackenzie (Fermilab)
- **Petascale Direct Numerical Simulations of Turbulent Channel Flow**
  - PI: Robert Moser (University of Texas)
- **Ab-initio Reaction Calculations for Carbon-12**
  - PI: Steven C. Pieper (Argonne National Laboratory)

Early Science Projects (con't)

- **NAMD - The Engine for Large-Scale Classical MD Simulations of Biomolecular Systems Based on a Polarizable Force Field**
  - PI: Benoit Roux (University of Chicago)
- **Global Simulation of Plasma Microturbulence at the Petascale and Beyond**
  - PI: William Tang (Princeton Plasma Physics Laboratory)
- **Multiscale Molecular Simulations at the Petascale**
  - PI: Gregory Voth (University of Chicago)

Early Tools Project

- Enabling Petascale Science on BG/Q: Tools, Libraries, Programming Models, & Other System Software (PI: Kalyan Kumaran)
  - Tools
    - PAPI, HPCToolkit, TAU, Scalasca, Open|SpeedShop, PerfSuite, FPMPI2
  - Debuggers
    - Allinea DDT, Rogue Wave TotalView
  - Libraries
    - Spiral, FFTW, Scalapack, BLAS , PETSc
    - Parallel I/O: MPI-IO, HDF5, Parallel NetCDF
    - Visualization, Chombo
  - Programming Models/Frameworks
    - Charm++, Coarray Fortran, GA Toolkit, MPI, UPC, GASnet
  - Other system software
    - Operating System Stacks

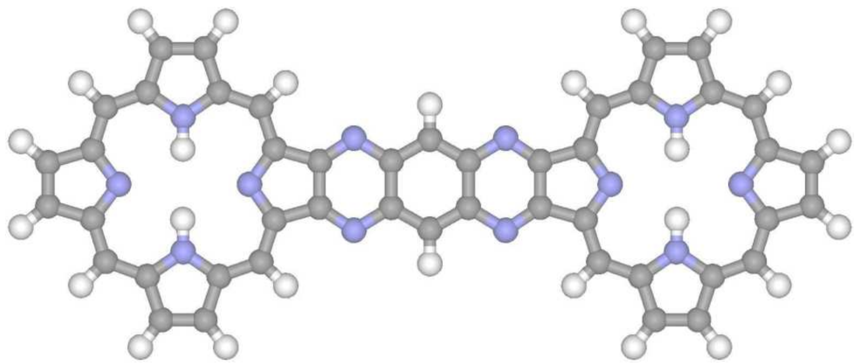
Leap To Petascale Workshops

- Annual multi-day workshops to focus on scaling and performance
  - Current INCITE , Discretionary projects
  - INCITE applicants to prepare proposals
- ALCF staff focus entirely on workshop
- External expertise for in-depth dives
  - Performance tools
  - Debuggers
  - IBM personnel

- L2P 2011
  - June 7-9, 2011
  - *Register by May 24*
  - http://workshops.alcf.anl.gov/petascale2k11/
- L2P 2010
  - Ex : Karniadakis (Brown) new INCITE project, Gordon Bell 2011 submission
  - Ex: Lin (GFDL) new ALCC
- L2P 2009
  - Significant progress on 8 projects
  - 7 INCITE proposals
  - Ex: Boldyrev new INCITE project
    - Scaled code from 4-32 racks
    - 40% performance improvement with ESSL implementation

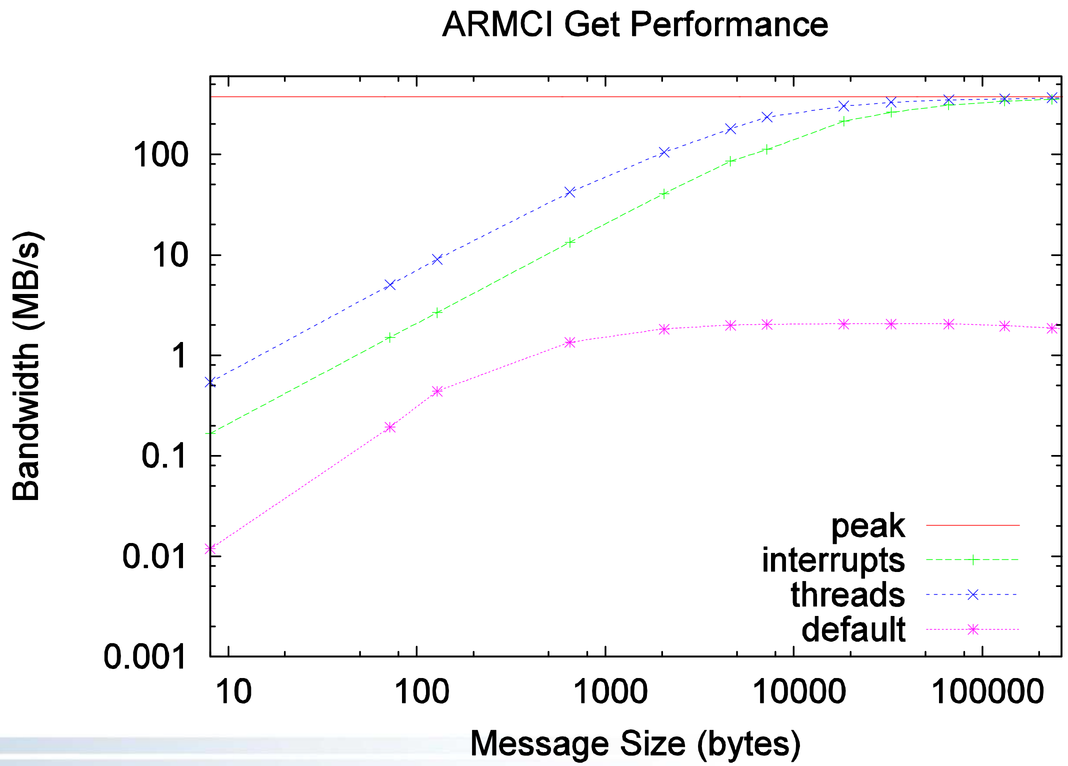
ARMCI – Jeff Hammond, ALCF

- NWChem computational chemistry package desired by multiple projects (INCITE and ALCC)
- NWChem relies upon Global Arrays and the ARMCI one-sided communication library, not just MPI
- ARMCI functional on Blue Gene/P but performance, scaling and stability not good in 2009
- Effective ARMCI bandwidth on BGP was 1% of what was possible due to undocumented disabling of DCMF interrupts in V1R3
- ARMCI had been untested by IBM on more than 1K nodes, preventing detection of non-scalable synchronization algorithms in ARMCI



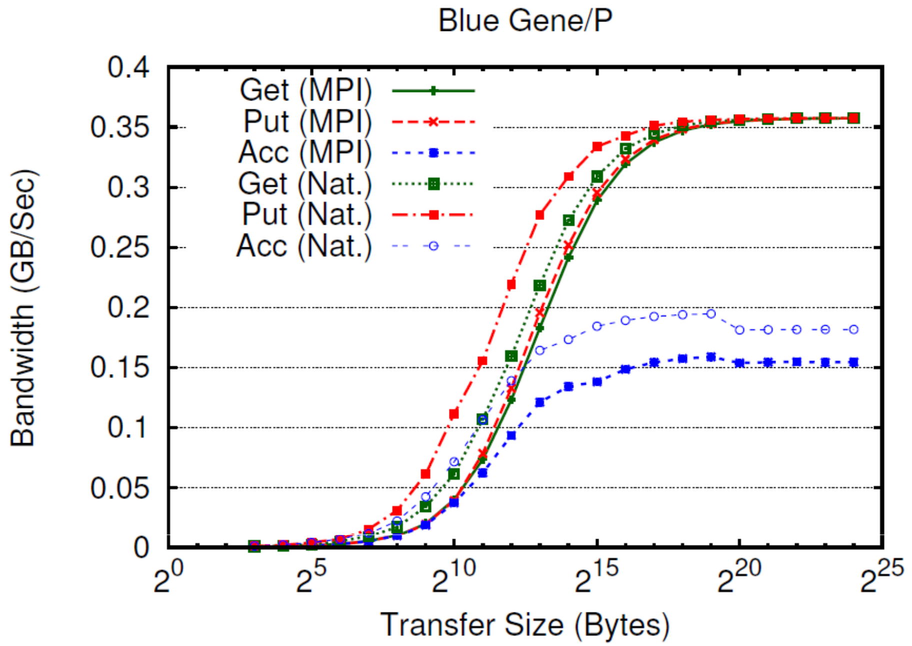
Performance Improvements ARMCI

- With help from IBM and PNNL, Jeff Hammond fixed ARMCI performance issues.
- Restored pre-V1R3 behavior by re-enabling interrupts and fixing MPI-compatibility issues.
- Implemented communication helper thread for NWChem, which runs in SMP mode because of memory requirements (1 comm thread + 1-3 compute threads



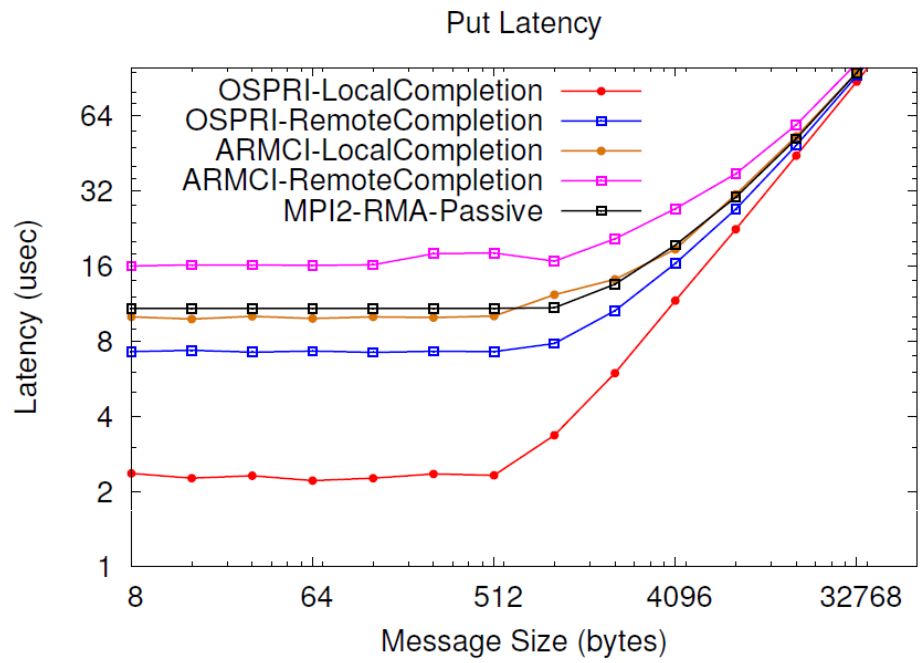
ARMCI-MPI: Portable ARMCI via MPI-2 RMA

- Jim Dinan of MCS implemented ARMCI over MPI-2 RMA, called ARMCI-MPI.
- MPI-2 RMA is implement on BGP; after a few bug fixes, it is a very satisfactory implementation.
- Performance with MPI is not as good as with DCMF, but it eliminates issues with direct use of DCMF.
- Assuming MPI-2 RMA works, ARMCI-MPI is Day 1 solution for NWChem on future IBM systems, e.g. Blue Gene/Q.
- See http://www.mcs.anl.gov/publications/paper_detail.php?id=1535 for ARMCI-MPI preprint.



Beyond ARMCI for One-sided Applications

- Jeff Hammond and Pavan Balaji designed OSPRI (One-Sided PRImitives) as successor to ARMCI.
- Design favors largest-scale systems, especially those with unordered networks.
- Relaxed consistency semantics (ordering) enable significantly better performance (see figure).
- Ivo Kabadshow and Holger Daschel of JSC use OSPRI predecessor to scale FMM to 300K of Jugene, which is not possible with MPI or ARMCI.

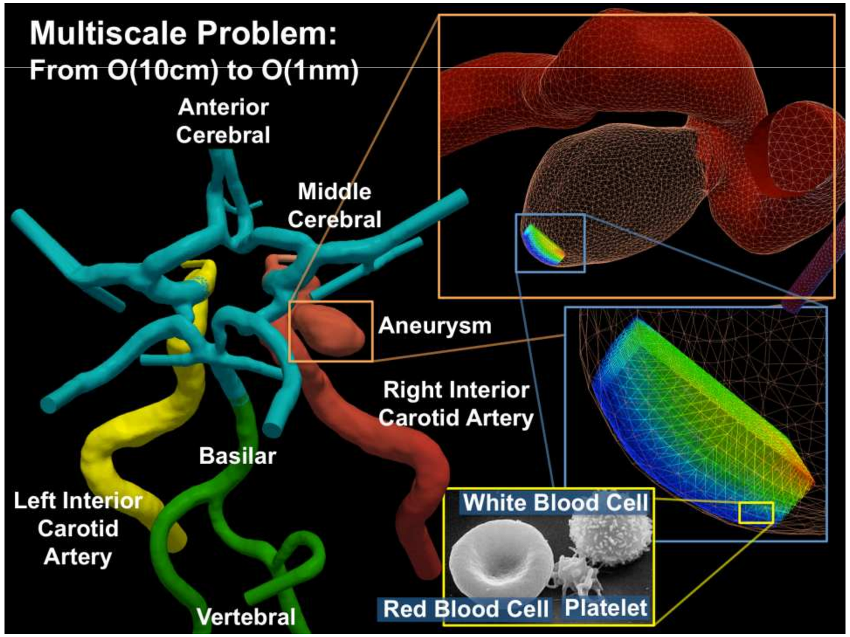


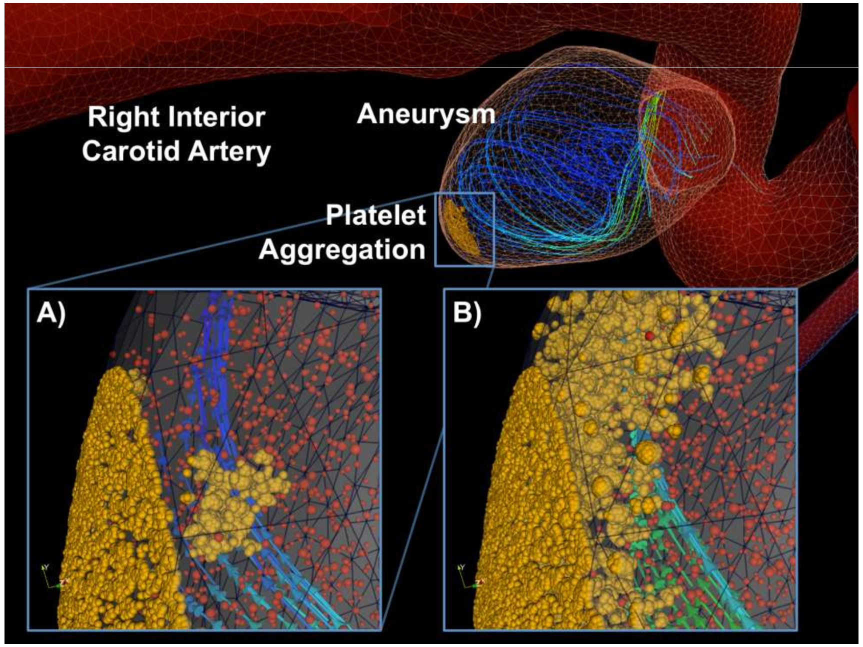
Multiscale Simulation in the Domain of Patient-specific Intracranial Arterial Tree Blood Flow (PI: George Karniadakis)

- **Goal:**
  - To perform a first-of-its-kind, multiscale simulation in the domain of patient-specific intracranial arterial tree blood flow.
- Code (NEKTAR-G) has two components:
  - NEKTAR
    - High-order spectral element code resolves large-scale dynamics
  - LAMMPS-DPD
    - Resolve mesoscale features
- Successfully integrated a solution of over 132,000 steps in a single, non-stop run on 32 compute racks of Blue Gene/P
- Frequent writes of 32GB to disk did not impact simulation

Multiscale Blood Flow (con't)

The computational domain consists of tens of major brain arteries and includes a relatively large aneurysm. The overall flow through the artery and the aneurysm as calculated by Nektar, as well as that within the subdomain calculated by LAMMPS-DPD, shown in detail in insets, along with platelet aggregation along the aneurysm wall.





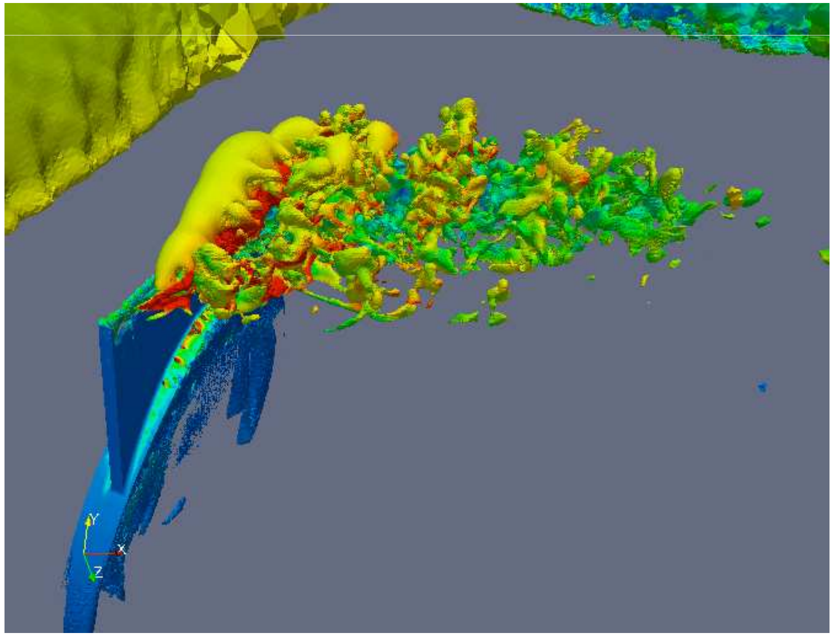
PHASTA (PI: Ken Jansen)

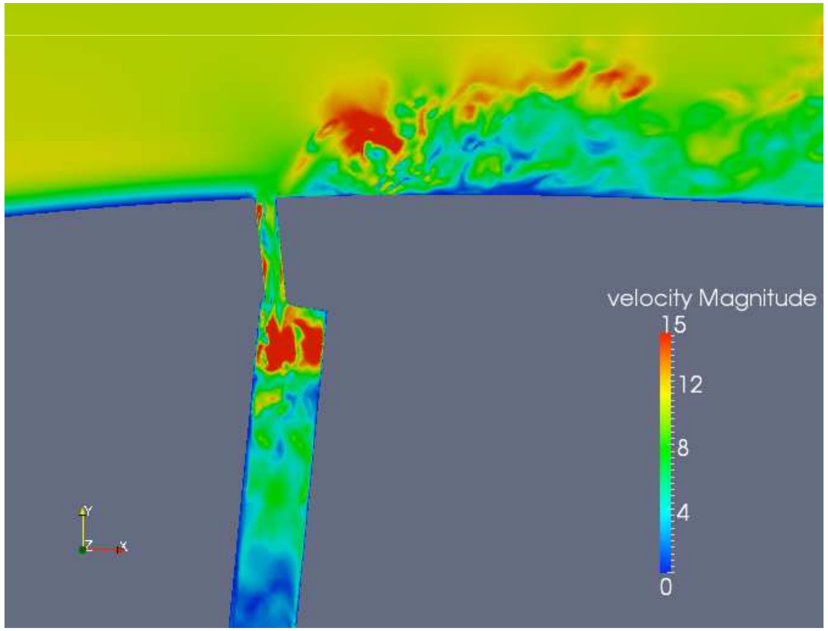
- Parallel, hierarchic (2nd-5th order accurate), adaptive, stabilized (finite element) transient, incompressible and compressible flow solver
- Can solve complex cases for which grid-independent solution can only be achieved through the efficient use of anisotropically adapted unstructured grids or meshes capable of maintaining high-quality boundary layer elements, and scalable performance on massively parallel computers.
- Scales to 288 thousand cores.
- GLEAN:
  - An MCS/ALCF-developed tool providing a flexible and extensible framework for simulation-time data analysis and I/O acceleration. GLEAN moves data out of the simulation application to dedicated staging nodes with as little overhead as possible.
- Collborative team (U Colorado, ALCF, Kitware) integrated latest GLEAN to collect data at large scale for PHASTA+GLEAN for three real-time visualization scenarios to determine frame rate and solver impact.

PHASTA (PI: Jansen)

The demonstration problem simulates flow control over a full 3D swept wing. Synthetic jets on the wing pulse at 1750Hz produce unsteady cross flow that can increase or decrease the lift, or even reattach a separated flow.

On the left is an isosurface of vertical velocity colored by magnitude of velocity and on the right is a cut plane through the synthetic jet (both on 3.3 billion element mesh). These are single frames taken from the real-time rendering of a live simulation.



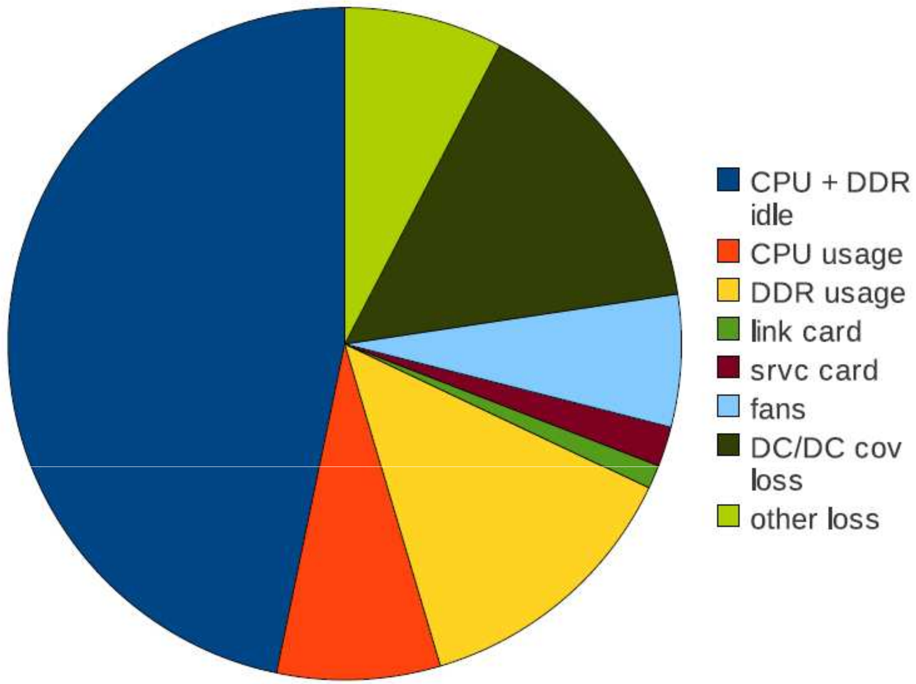


Power Consumption and Power Management on BG/P (William Scullin and Chenjie Yu)

- Power consumption has emerged as the a critical factor in both individual node architecture and overall system designs.
- Blue Gene at the top of ”green computing” list but yet the ANL BG/P costs more than one million dollars/year in electricity
- Implications for Exascale
- In this project:
  - Utilized the existing Environment Monitoring mechanisms in BG/P
  - Experimented on a set of test programs stressing different parts of the system, to break down the power consumption to different components.
  - Also explored ways to reduce BG/P power consumption by using built-in throttling mechanisms and CPU power saving mode in ZeptoOS

Power Consumption and Management (con't)

- Breakdown of power use by Lattice QCD (*right*)
- Pro-active power management (*below*)
  - Processor throttling
    - No significant drop
  - Memory throttling
    - Up to 32% lower



Benchmarks	DFP	PctChg	L1-stress	PctChg	L3-stress	PctChg	Sleep	PctChg
BULK	23,061.06	1.04%	22,803.60	0.23%	22,397.48	16.48%	20,177.66	3.56%
NODECARD	15,556.78	3.45%	16,009.12	0.61%	15,405.25	20.20%	13,799.94	0.16%
Benchmarks	m-stress	PctChg	QCD	PctChg	AllToAll	PctChg	Empty	PctChg
BULK	22,815.63	26.62%	22,604.87	23.42%	21,988.55	3.63%	N/A	N/A
NODECARD	15,301.53	32.43%	15,578.16	22.33%	15,160.14	7.13%	N/A	N/A

Large-Scale System Monitoring Workshop
Argonne Leadership Computing Facility
May 24-26, 2010

Hosted by Bill Allcock, ALCF Director of Operations and Randal Rheinheimer, Deputy Group Leader for HPC Support at LANL:

- 19 attendees from ANL, LANL, IU, LBNL, SNL, LLNL, KAUST, INL, and NCSA.
- Day 1: Institutions gave overviews of their systems and monitoring, noting if their current solutions were good or if improvements were needed.
- Day 2: The group worked to define “monitoring” and discussed potential issues with increased scale, plus what precipitates a move towards common monitoring infrastructure (money, resources, cultural change, etc.)
- Action Items:
  1. An “exascale monitoring” BOF at SC10 to broaden participation
  2. A mailing list for asking questions of the group
  3. A wiki for gathering “monitoring best practices”
  4. An “exascale monitoring” white paper

In Summary

- ALCF BG/Q Mira is on the way
- The Early Science Program will bridge the gap from BG/L to BG/P
- Deadlines
  - Leap to Petascale Workshop register by May 24
  - INCITE 2012 deadline 6/30/2011
- http://www.alcf.anl.gov